# Mixed driven Refinement design of Multidimensional models based on Agglomerative Hierarchical Clustering

Lucile Sautot, Sandro Bimonte, Ludovic Journaux, Bruno Faivre

# Mixed driven Refinement design of Multidimensional models based on Agglomerative Hierarchical Clustering

Lucile Sautot[1,2], Sandro Bimonte[3], Ludovic Journaux[4] and Bruno Faivre[1]
[1]*Biogéosciences Laboratory, University of Burgundy, 6 boulevard Gabriel, Dijon, France*
[2]*AgroParisTech,19 rue du Main, Paris, France*
[3]*IRSTEA Centre de Clermont-Ferrand, 9 avenue Blaise Pascal, Aubière, France*
[4]*LE2I, University of Burgundy, allé Alain Savary, Dijon, France*
{*lucile.sautot, bruno.faivre*}*@u-bourgogne.fr, sandro.bimonte@irstea.fr, ludovic.journaux@agrosupdijon.fr*

Keywords: Multidimensional design, Data Warehouse, OLAP, Data Mining

Abstract: Data warehouses (DW) and OLAP systems are business intelligence technologies allowing the on-line analysis of huge volume of data according to users' needs. The success of DW projects essentially depends on the design phase where functional requirements meet data sources (mixed design methodology) (Phipps and Davis, 2002). However, when dealing with complex applications existing design methodologies seem inefficient since decision-makers define functional requirements that cannot be deduced from data sources (data driven approach) and/or they have not sufficient application domain knowledge (user driven approach) (Sautot et al., 2014b). Therefore, in this paper we propose a new mixed refinement design methodology where the classical data-driven approach is enhanced with data mining to create new dimensions hierarchies. A tool implementing our approach is also presented to validate our theoretical proposal.

## 1 INTRODUCTION

Data warehouses (DW) and OLAP systems are business intelligence technologies allowing the on-line analysis of huge volume of data. Warehoused data is organized according to the multidimensional model that defines the concepts of dimensions and facts. Dimensions represent analysis axes and they are organized in hierarchies. Facts are the analysis subjects and they are described by numerical indicators called measures. Warehoused data are then explored and aggregated using OLAP operators (e.g. Roll-up, Slice, etc.) (Kimball, 1996).

The success of DW projects essentially depends on the design phase where functional requirements meet data sources (Phipps and Davis, 2002). Three main methodologies have been developed: user-driven, datadriven and mixed (Romero and Abello, 2009). User-driven approach puts decision-makers at the center of the design phase by providing them tools to define the multidimensional model exclusively according to their analysis needs. Usually, data driven methodology proposals deduce the multidimensional model from structured and semistructured (Mahboubi et al., 2009; Jensen et al., 2004) data sources exploiting metadata (e.g. foreign keys) and some empirical values. Finally, mixed approaches fusion the two previous described methods.

Hierarchies are crucial structures in DW since they allow aggregation of measures in order to provide a global and general analytic view of warehoused data. For that reasons, some works investigate definition of hierarchies by means of Data Mining (DM) algorithms (Favre et al., 2006; Sautot et al., 2014b). However, this design step is applied once the multidimensional model has been defined, and it takes into account only members of one dimension.

From our point of view, these methodologies present an important limitation since in real DW projects often those DM algorithms need data of different dimensions and facts. Thus, in this paper we present a framework for a mixed design of multidimensional models by integrating DM algorithms in a classical data driven-approach. This allows defining hierarchical structures, according to decisional users' requirements, that cannot be deduced by classical datadriven methods. This hierarchical organization of dimensional data is translated in a complex multi-factual multidimensional model in order to represent as well as possible semantic of data sources.

The paper is organized in the following way: Section 2 introduces related work; a retail case study and the motivation are presented in Section 3; our design method is detailed in Section 4 and its implementation is shown on Section 5.

## 2 RELATED WORK

Three types of approaches can be used to design a data warehouse: (i) Methods based on user specifications, or demand-driven approaches; (ii) Methods based on available data, or data-driven approaches; (iii) Mixed methods, or hybrid approaches. For example, (Jovanovic et al., 2012) is an iterative demand-driven method where at each iteration, the system searches for the best data corresponding with the information required by the user in terms of dimensions or facts. Moreover, several other have proposed systems based on hybrid approach such as (Romero and Abello, 2010) that propose to express functional requirements using SQL queries.

Relational data driven approaches deduce multidimensional structures (facts and dimensions) from conceptual (Phipps and Davis, 2002) and/or logical models (Carme et al., 2010; Jensen et al., 2004). In particular some works investigate automatic discovering facts using some heuristics (Carme et al., 2010). About dimensions some works propose using logical database metadata such as foreign keys (Jensen et al., 2004) or some heuristics.

Other works use more complex algorithm to identify dimensions hierarchies. (Nguyen and Tjoa, 2000) propose a system to dynamically build hierarchies based on data from Twitter (Nguyen and Tjoa, 2000). (Messaoud et al., 2004) present a new OLAP operator named OPAC that allows to aggregate facts that refer to complex objects, such as images. This operator is based on hierarchical clustering algorithm. (Favre et al., 2006) provide a framework for automatic defining hierarchies according to user rules. In order to personalize the multidimensional schema, (Bentayeb, 2008) propose to create new levels in a hierarchy with the K-means algorithm. (Leonhardi et al., 2010) propose to increase the OLAP cube exploration functionalities by providing the user data mining algorithms to analyze data. (Ceci et al., 2011) use a hierarchical clustering to integrate continuous variables as dimensions in an OLAP schema. In the same line, (Sautot et al., 2014b) propose using Agglomerative Clustering for designing hierarchies, and the integration in a rapid prototyping methodology is presented in (Sautot et al., 2014a). However, all existing works define hierarchies using only either dimensional data (i.e. attributes of dimension members) or factual data (i.e. measures) (see Table 1). But, in a constellation schema, a dimension can be enriched with a hierarchy created by using other dimensions and facts. It means that the creation of a new hierarchy can involved a refinement of facts and dimensions in the entier constellation schema. We detail this issue in the following section, using a real application case from bird biodiversity.

| | | Data sources | | |
|---|---|---|---|---|
| | | Star schema | | Constellation schema |
| | | Facts | One dimension | Facts and dimensions |
| Algorithm | K-means | (Bentayeb, 2008) | (Bentayeb, 2008) | |
| | Hierarchical classification | (Ceci et al., 2011) (Sautot et al., 2014a) (Sautot et al., 2014b) | (Messaoud et al., 2004) | Our proposal |
| | Other | (Favre et al., 2006) (Nguyen and Tjoa, 2000) | (Leonhardi et al., 2010) | |

Table 1: Summary of literature review related to automatic hierarchy building

## 3 MOTIVATION

In order to describe motivation of our new DW design methodology we present in this section a real case study concerning the bird biodiversity analysis (Sautot et al., 2014b). This dataset has been collected to analyze spatio-temporal changes in bird populations along the Loire River (France) and to identify local and global environmental factors that can explain these changes. Data sources are stored in a relational database (PostGIS). Applying the data driven algorithm proposed in (Romero and Abello, 2010) we obtain the constellation schema depicted in Figure 1, which presents two facts as described in the following. Abundances is one fact, and can be analyzed according to three dimensions (an instance is shown on Table 3): (i) the species dimension, which stores species names and attributes, (ii) the time di-

mension, which corresponds to the census years and (iii) the spatial dimension, which describes census points along the river. Using this model decision-maker can answer to queries like: *"What is the total of birds per year and census point?"* or *"What is the total of birds per year and altitude?"*. To complete bird census, the landscape and the river are described around each census point. Environment descriptions are represented by another fact, which is associated to the time dimension and the spatial dimension. With this model, it is possible to describe census points, for example a possible OLAP query is *"What is the percentage of forest per census point in 2012?"*.

Note that descriptions of census points that are not dependent from time, such as altitude and geology, are used as spatial dimension levels, while other attributes are represented as measures of another fact (e.g. percentage of forest). Unfortunately, abundances for a specie have not meaning if not related to environmental data of census points. In this situation a drill-across operation is not adequate since it will hide the species dimension. Indeed, with the drill-across operators facts are joined only on common dimensions. Moreover, the multidimensional model of Figure 1 does not make possible to provide the decision-makers with OLAP queries aggregating abundance by classes of environmental variable (30% of forest, 50% of water, etc.), for example *"What is the total of birds per year and group of census point with 30% of forest?"* or *"What is the total of birds per year and group of census point with 50% of water?"*, since environmental parameters do not appear as levels, but as measures, prohibiting group-by queries.

Therefore, in our case study, decision-makers need for a new design method that group census points (dimensional data) by environmental parameters (factual data) and year (dimensional data).

The multidimensional model allowing correct OLAP analysis should be the one shown on Figure 2 (Miquel et al., 2002b). This multidimensional schema presents only one fact and the spatial dimension is enriched with some levels representing group of environmental parameters for each year. Indeed, environmental parameters for census points in 2001 can be different from ones of 2002 implying that the same census point is not grouped in the same level on two different years as shown on Table 3.

For example, data describing agricultural activities around the census points, are available only for the 2002 census campaign. Therefore, it is important to take into this different classification when navigating on the temporal dimension during an OLAP analysis session. For example, the query *"What is the total of birds in 2002 and in census points with the same environmental parameters?"* has to use the environment type 2002 level, and *"What is the total of birds in 2011 and in census points with the same environmental parameters?"* has to use the environment type 2011 level. For example an OLAP query using the environment type 2002 level and the temporal member 2011 is not coherent since it associates the number of birds on 2011 in the past geographical-environmental configuration of 2002, leading to erroneous interpretation.

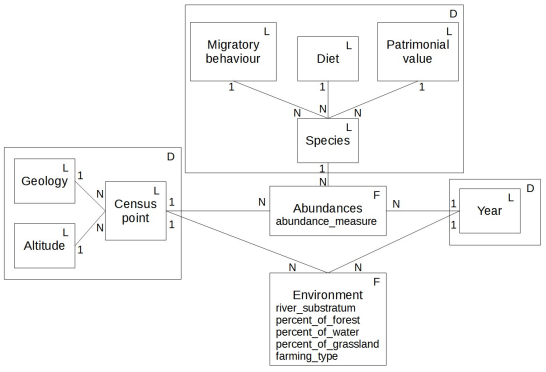


Figure 1: Bird biodiversity case study: Data-driven constellation schema

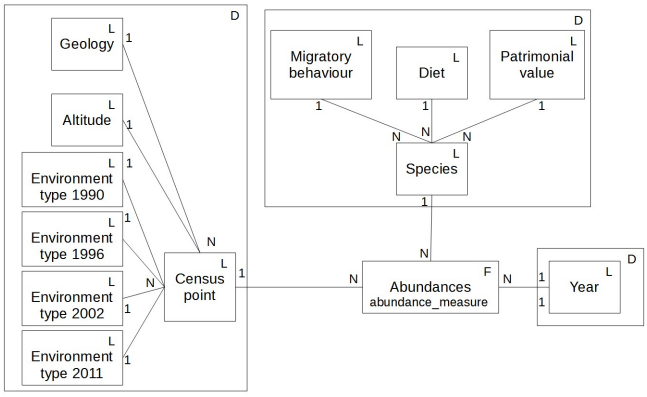


Figure 2: Bird biodiversity case study:manually driven multi-version schema

Table 2: Factual data of "Environments" node

| Years | Census Points | Agencies | Percent of Forest | Percent of Grass-land |
|---|---|---|---|---|
| 2002 | 1 | LE2I | 0.176 | 0.250 |
| 2002 | 1 | ONEMA | 0.356 | 0.261 |
| 2002 | 2 | LE2I | 0.311 | 0.420 |
| 2002 | 2 | ONEMA | 0.255 | 0.574 |
| 2011 | 1 | LE2I | 0.189 | 0.278 |
| 2011 | 1 | ONEMA | 0.241 | 0.385 |
| 2011 | 2 | LE2I | 0.322 | 0.568 |
| 2011 | 2 | ONEMA | 0.257 | 0.575 |

Table 3: Factual data of "Abundances" node

| Years | Census points | Species | Abundance |
|---|---|---|---|
| 2002 | 1 | Yellowhammer | 1.5 |
| 2002 | 1 | Coal Tit | 0.5 |
| 2002 | 2 | Yellowhammer | 1.5 |
| 2002 | 2 | Coal Tit | 0 |
| 2011 | 1 | Yellowhammer | 1 |
| 2011 | 1 | Coal Tit | 3 |
| 2011 | 2 | Yellowhammer | 1 |
| 2011 | 2 | Coal Tit | 2 |

# 4 OUR PROPOSAL

In this section we introduce our framework for the refinement of multidimensional in a mixed approach. The main idea of our proposal is using an existing data driven methodology in a first step. Then, in our new design step, we collect user needs about hierarchies that are not been deduced in the multidimensional schema by means of the functional dependencies. These users' needs are expressed in the form of facts existing in the constellation multidimensional model. In particular, the main idea is to provide an algorithm that transforms the constellation multidimensional schema by eliminating a fact node and integrating factual data in an associated dimension used for creating new levels.

To perform this algorithm, we translate the multidimensional model in a multidimensional graph.

In the following section we describe the multidimensional graph definitions (Section 4.1), the main algorithm is detailed in Section 4.2 and the calculation of new versioned hierarchies is explained in Section 4.3.

## 4.1 Preliminaries

In this subsection, we present some preliminary definitions.

We represent a multidimensional model using a graph.

*Definition 1.* **Multidimensional graph.**

A multidimensional graph is a directed graph $M_G =< D, F, A >$ with:

$D = \{d_1, ..., d_m\}$, dimensional nodes, which represent dimensions.

$F = \{f_1, ..., f_n\}$, fact nodes representing facts.

$A = \{a_1, ..., a_p\} \mid \forall i \in [\![1, p]\!], a_i = (f_j, d_k)$, with $j \in [\![1, n]\!]$ and $k \in [\![1, m]\!]$, are arcs[1], meaning that arcs are only directed from a fact node to a dimensional node.

Moreover, $M_G$ contains no alone node, isolated of another node, but can contain possibly disconnected sets of nodes if each sub-graph must contain at least one fact node.

*Example.* An example of multidimensional graph is shown on Figure 3. "Species" dimension, "Census points" dimension, "Years" dimension, "Abundances" fact and "Environments" fact are described in previous sections. "Sources" dimension represents agencies, which collect data. "Budget" fact represents the funds allowed by each agency for each year to collecting data.

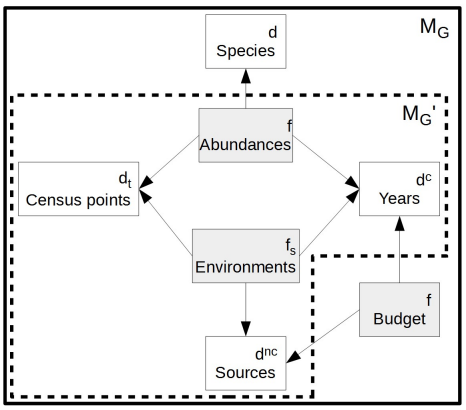


Figure 3: Multidimensional graph $M_G$

In our approach decision-maker want to enrich a dimension with some new hierarchies using some factual data. That dimension is called Target dimension

*Definition 2.* **Target dimension.**

The target dimension $d_t$ of a multidimensional graph $M_G$ is a dimension such as:

$$d_t \in D \mid \exists (f_1, d_t), ..., (f_u, d_t) \; with \; u \in [\![2, n]\!]$$

[1]In this paper, the notation $(f_i, d_j)$ represents the arc from fact node $f_i$ to dimensional node $d_j$.

This means that dt is associated at least to two facts since one has to be removed and used to create its new levels.

*Example.* An example of possible target dimension is the "census point" dimension (Figure 3).

Let us now formalize the fact node that is used to create levels.

*Definition 3.* **Source node.**

The source node of a $M_G$ with a target dimension $d_t$ is a fact node $f_s \in \{f_1, ..., f_u\}$.

*Example.* With "census point" dimension as target node, an example of possible source node is the fact node "Environments".

As we have said before our algorithm removes the source node from the graph. Therefore, a part of the structure of the graph is changed. Note that only nodes related to the source nodes are affected. We define this sub-graph in the following way

*Definition 4.* **Source-target multidimensional sub-graph.**

Let $M_G$ a multidimensional graph with a target dimension $d_t$ and a source node $f_s$ then the Source-target multidimensional sub-graph $M'_G$ is a multidimensional graph such as: $M'_G =< D', F', A' >$ with:

$F' = \{f_i \in F \mid \exists(f_i, d_t)\}$

$D' = \{d_i \in D \mid \exists(f_s, d_i)\}$

$A' = \{(f_i, d_j) \mid f_i \in F', d_j \in D'\}$

$M'_G$ contains thereby only fact nodes linked to $d_t$ and dimensional nodes linked to $f_s$. In $M'_G$, all fact nodes are so linked to at least one dimensional node and all dimensional nodes are so linked to at least one fact node. There is no isolated node in this sub-graph. $M'_G$ is so a well-formed multidimensional graph.

*Example.* An example of Source-target multidimensional sub-graph using the previous example is shown on Figure 3.

In order to formalize inputs of the agglomerative hierarchical clustering algorithm used for the creation of levels of the target dimension, we formalize factual data aggregated to a set of dimensions levels using the definition of instance fact node.

*Definition 5.* **Instance fact node.**

Let $M_G$ a multidimensional graph. Let $m_i$ a member of the dimension $d_i$. Then the instance fact node $I(f, d_1.m_1, ..., d_n.m_n)$ is the set of tuples representing facts of $f$ aggregated to the dimensions members $d_1.m_1$, ..., $d_n.m_n$.

*Example.* Let, Table 2 representing the instance fact node for the node "Environments", then Table 4 represents facts aggregated to the All member of the "Agencies" dimension:

(I("Environments", "Agencies.ALL", "Years.1990", "Census_points.*"))[2]

## 4.2 Algorithm

In this section we provide details and formalize our approach.

Removing a fact node from the multidimensional graph implies its redefinition. Thus, the main idea is in a first step to work on the source-target multidimensional graph exclusively, transform this sub-graph adding levels to the target dimension and removing the source node, and then finally re-integrate the new sub-graph in the rest of original multidimensional graph.

Removing the source node implies to handle its associated dimensions. It is possible to distinguish three types of dimensions:

- The target dimension that will rest in the transformed sub-graph,
- the Non Context dimensions $D_{nc}$, and
- the Context dimensions $D_c$.

The Non context dimensions $D_{nc}$ are dimensions that are only associated to the source node fact. In order to remove one dimension it is possible to provide a classical Dice operator, which consists in aggregating fact data to the top dimension member. Let us note that in order to avoid summarazability problems (aggregation cannot be reused) (Lenz and Thalheim, 2009), in our approach we allow using only distributive and algebraic aggregation functions for the Dice operator.

*Example.* An example of Non contextual dimension is the "Agencies" node. In Table 4 is shown an example of the Dice operator on the Agencies dimension, which is a Non contextual dimension.

Table 4: Factual data of "Environments" node aggregated on "Agencies"

| Years | Census Points | Percent of Forest | Percent of Grass-land |
|---|---|---|---|
| 2002 | 1 | 0.266 | 0.256 |
| 2002 | 2 | 0.283 | 0.497 |
| 2011 | 1 | 0.215 | 0.332 |
| 2011 | 2 | 0.290 | 0.572 |

Formally,

*Definition 6.* **Non contextual dimension.**

[2] '*' means 'all members of the dimension'

Let Source-target multidimensional sub-graph $M'_G =< D', F', A' >$, then the set of non contextual dimension $D_{nc}$ is

$$D_{nc} = \{d_1^{nc}, ..., d_v^{nc}\} \subset D' \mid \forall i \in [\![1, v]\!] \exists! (d_i^{nc}, f_j) \mid f_j \in F'$$

Note that in the previous formula, all dimensional nodes in $D_{nc}$ are only linked to $f_s$. Indeed, all dimensional nodes in $M'_G$ are linked to $f_s$ and dimensional nodes in $D_{nc}$ are linked to one (and only one) dimensional node.

The Context dimensions $D_c$ are dimensions in $M'_G$ that are associated to $f_s$ and another fact node $f$. With the future refined graph, users analyze facts in $f$ according to $d_t$. But, data used for calculating new hierarchies in $d_t$ come from $f_s$ and are thereby dependent of dimensions in $D_c$. Therefore, we need to ensure that data used to create the hierarchy are coherent with data consulted by the user during their OLAP analysis. With this in mind, we offer a system that calculates hierarchies according a context, this context defining with $D_c$.

Formally,

*Definition 7.* **Contextual dimension.**
Let Source-target multidimensional sub-graph $M'_G$, then the set of contextual dimension $D_c$ is

$$D_c \subset D' \mid D_c = D' - (D_{nc} \cup \{d_t\})$$

*Example.* An example of contextual dimension is the "Years" node. On Table 3, we present data from "Abundances" node: data are dependent of "Years" dimensional node.

Once we have defined non context and contex dimensions let us provide our algorithm supposing that we have only one context dimension.

The input of this algorithm is the multidimensional graph $M_G$ presented on Figure 3.

***Begin of the Refinement Algorithm***

1. *Identify the Source-target multidimensional sub-graph $M'_G$.*
2. *Calculate a hierarchy for each instance of each context. This part of the algorithm is detailed in particular in the section 4.3.*
3. *Remove $f_s$ from $M_G$.*
4. *Remove isolated nodes. The isolated nodes can be only dimensional nodes linked to $f_s$. Then $M_G$ is well formed.*

***End of the Refinement Algorithm***

The output of this algorithm is a multidimensional graph, presented on Figure 4. We note that $f_s$ has been removed and there are new hierarchies in the "census points" node. Moreover, $M_G$ remains a well-formed multidimensional graph and can be also implmented in a ROLAP architecture.

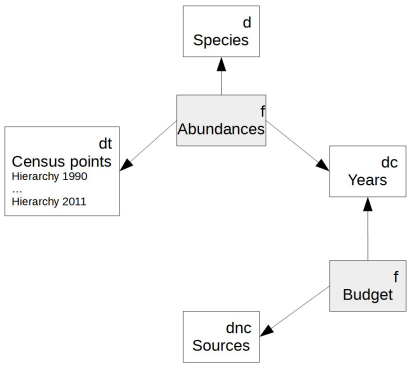


Figure 4: Refined multidimensional graph $M_G$

## 4.3 Automatic creation of hierarchies

In this section we describe how the is applied to create new levels of the target dimension.

A complete methodology to create new hierarchies in a multidimensional model with Hierarchical Agglomerative Clustering is presented in (Sautot et al., 2014b). The main idea of this methodology is to build a new hierarchy into a dimension by using data, which describe items at the lowest level of the hierarchy. In our case, items are census points and description data are factual data. We suggest to use the Hierarchical Agglomerative Clustering, due to the similarity between the output of the Hierarchical Agglomerative Clustering and a hierarchy into an OLAP dimension (Messaoud et al., 2004).

Main steps of this algorithm are: (1) Calculation of distances between individuals; (2) Choice of the two nearest individuals. (3) Aggregation of the two nearest individuals in a cluster. The cluster is considered an individual. (4) Go back to the step 1 and loop while there is more than one individual.

In our approach the clustering (AHC) takes as inputs the instance of the source node $f_s$ evaluated on each member of the context dimension and dicing it non context dimensions.

Formally, the step 2 of our algorithm is the following:

***Begin of the Hierarchy Builder Algorithm***
*for each member$_i$ of $d_c$*
*. create a new hierarchy of $d_t$*
*. AHC(I($f_s$,$d_1^{nc}$.ALL,... ,$d_v^{nc}$ . ALL, $d^c$.member$_i$, $d_t$.*))*
***End of the Hierarchy Builder Algorithm***

An example is presented on Figure 5. We note that two hierarchies for the spatial dimension have been

created for years 2002 and 2011.

| Years | Census Points | Forest | Grassland |
|---|---|---|---|
| 2002 | 1 | 17.55% | 30.29% |
| 2002 | 2 | 31.12% | 47.41% |
| 2002 | 3 | 41.29% | 58.71% |
| 2002 | 4 | 0.00% | 0.00% |
| 2002 | 5 | 15.31% | 84.69% |
| 2011 | 1 | 35.62% | 62.16% |
| 2011 | 2 | 22.27% | 74.85% |
| 2011 | 3 | 58.50% | 13.72% |
| 2011 | 4 | 47.88% | 21.51% |
| 2011 | 5 | 37.06% | 37.96% |

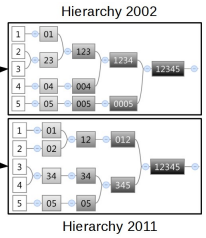


Figure 5: Contextual hierarchies of census points

# 5 VALIDATION AND EXPERIMENTS

In this section we present the implementation our proposal. A semantic and performance evaluations are detailed in Sec 5.1 and 5.2 respectively.

The refinement tool implements our algorithm using Matlab. It allows defining graph using a simple visual interface as shown on Figure 7. The considered multidimensional graph is presented on the top part of the visual interface. On the bottom one, the algorithm ask inputs to users in a command window.

## 5.1 Semantic evaluation

In this section, we describe the added-value of our methodology from a design point of view (i.e. does the refinement methodology corresponds to decision-makers needs?). For that goal two we have investigated two aspects: 1) Do dimensions and facts created using our methodology correspond to decision-makers analysis needs?; 2) Do hierarchies created using our methodology improve analysis capabilities?

Therefore have decided to compare the result of our methodology with with one proposed in (Miquel et al., 2002a). Indeed, (Miquel et al., 2002a) propose a manually method to obtain a multi-version multidimensional schema,and when the time dimension is chosen as the context dimension our approach results a multi-version multidimensional schema. The result of this validation shows that the multidimensional schema produced with the manual methodology and our automatic methodology are equal.

Moreover, in order to validate the semantic correctness of using AHC for hierarchies definition, we have asked to ecologists of the project to choice between a spatial dimension with only one level, and a spatial dimension with a hierarchy created using AHC. When the number of created levels is not superior to 5, decision-makers prefer having hierarchies, since they can reveal interesting pattern such as agricultural profiles of census points. For example, data in the "Environments" fact table contains data that describe agriculture policies around each census point at each year. The data clustering according to these data can classify census points and allows decision-makers analyzing impact of agricultural practices on bird biodiversity. For example, decision-makers can analyze biodiversity according to agricultural forest and grassland parameters of census points, by using this simple OLAP query: "What is the biodiversity value per group of census points (first level of the hierarchy obtained with clustering) in 2002 and 2003?". This query can reveal that for the same year, for example 2002, biodiversity is very affected by agricultural parameters since the aggregated biodiversity value for each group of census point is different.

## 5.2 Performance evaluation

In this section, we test time performance of our methodology in order to validate its feasibility from a project deployment process point of view.

In particular we study time performance related to: 1) refinement algorithm for facts and dimension design, and 2) hierarchy creation using AHC.

In order to test the first point, we have created a set of 200 simulated constellation schema using from 2 to 100 dimensions, since real usable multidimensional schema presents maximum between 3 and 10 dimensions (Kimball, 1996). Finally, the worst time execution is 15.23 s. The average execution time is equal to 11.7 s with a standard deviation equal to 1.17 s. These performances are satisfactory for are good for an off-line design phase.

In this paragraph, we study time performances of the AHC algorithm. In this paragraph, "classified items" are census points (which are members of the "census points" dimension, the target dimension) and "attributes" are aggregated facts from the "Environments" fact node (which is the source fact node). The AHC algorithm has been also implemented in Matlab and its performance has been also tested. Using our case study data, we perform 2090 tests, with a number of classified items (source node instances-Enverinments facts) between 10 and 190, and a number of attributes (source node attributes-Enverinments fact measures) between 10 and 100, and the average calculation time is equal to 0.072 s, with a standard deviation equal to 0.002 s. To complete our evaluation, we simulate a data set with 10,000 classified items and 150 attributes. In this case, the AHC calculates a hierarchy in 147.36 s, with a standard deviation equal to 4.03 , with a maximal calculation time equal

to 214 s. All time performances are shown on Figure 6. This calculation time (approximately four minutes) is efficient for an off-line design phase.

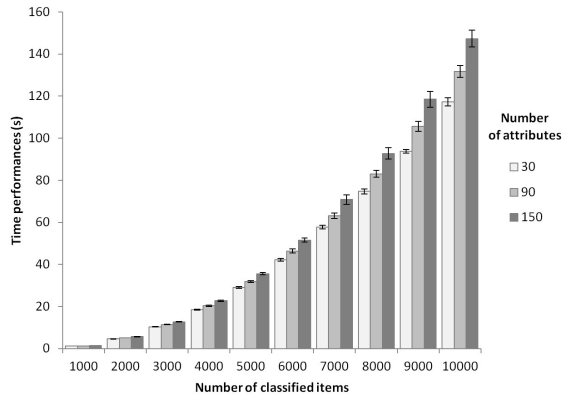


Figure 6: Execution times according the number of attributes and classified items

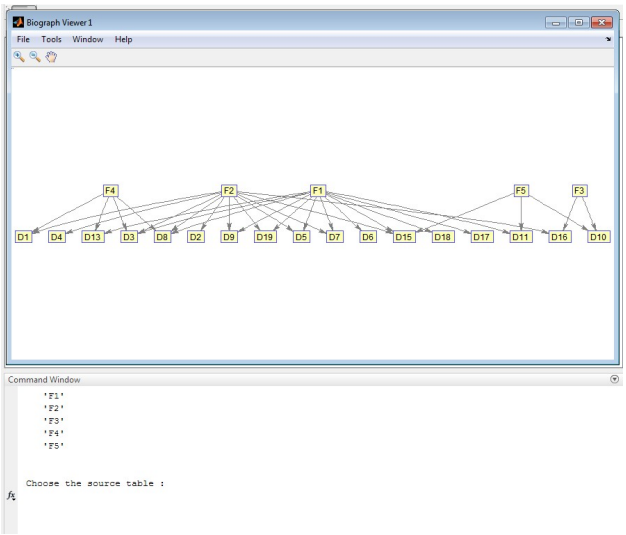


Figure 7: Visual interface of the refinement tool

# 6 CONCLUSION AND FUTURE WORK

Design data warehouses system is a complex and crucial task depending on available data sources and decisional requirements. Existing work do not exploit the semantics of data to automatically create complex hierarchies. Thus in this paper, we present a mixed multidimensional refinement methodology, that transform constellation schema to define hierarchy level using a hierarchical clustering algorithm. Our refinement methodology enriches a dimension with factual data, and considers the context of factual data. We present also the implementation of our method in a ROLAP architecture.

We perform the proposed methodology on a real application case from bird biodiversity. We have noted that actual automatic multidimensional design methodologies cannot produce a multidimensional schema, which covers all decision-maker needs due to the data complexity. Our methodology offers a solution to enrich dimensions with factual data and, by this way, to refine the multidimensional schema.

Our ongoing work is the extension of our methodology to simplify and reduce the number of created levels, using other DM algorithms such as SVM, etc., in order to provide decision-makers with easy OLAP exploration analysis and its implementation in a ROLAP architecture.

Moreover, we are also working to integrate our approach in the rapid prototyping methodology proposed in (Sautot et al., 2014a), and extending to help decision-makers and DW experts choose the right DM algorithms and parameters of the refinement algorithm (source node, contextual dimensions, etc.). Future work concerns the usage of the formal evaluation framework Goal Question Metric (Briand et al., 2002) to evaluate our methodology.

# ACKNOWLEDGMENTS

Data acquisition received financial support from the FEDER Loire, Etablissement Public Loire, DREAL de Bassin Centre, the Région Bourgogne (PARI, Projet Agrale 5) and the French Ministry of Agriculture. We also thank heartily Pr. John Aldo Lee, from the Catholic University of Leuven, for his help.

# REFERENCES

Bentayeb, F. (2008). K-means based approach for olap dimension updates. In *10th International Conference on Enterprise Information Systems (ICEIS)*, pages 531–534.

Briand, L. C., Morasca, S., and Basili, V. R. (2002). An operational process for goal-driven definition of measures. *IEEE Trans. Software Eng.*, 28(12):1106–1125.

Carme, A., Mazon, J.-N., and Rizzi, S. (2010). A model-driven heuristic approach for detecting multidimensional facts in relational data sources. In Pedersen, T., Mohania, M., and Tjoa, A. M., editors, *Proceedings of 12th International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, volume LNCS 6263, pages 13–24.

Ceci, M., Cuzzocrea, A., and Malerba, D. (2011). Olap over continuous domains via density-based hierarchical clustering. In *15th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2011)*, volume 2, pages 559–570.

Favre, C., Bentayeb, F., and Boussaid, O. (2006). A knowledge-driven data warehouse model for analysis evolution. *Frontiers in Artificial Intelligence and Applications*, 143:271.

Jensen, M. R., Holmgren, T., and Torben (2004). Discovering multidimensional structure in relational data. In *Data Warehousing and Knowledge Discovery: 6th International Conference (DaWaK)*.

Jovanovic, P., Romero, O., Simitsis, A., and Abelló, A. (2012). Ore: An iterative approach to the design and evolution of multi-dimensional schemas. In *Proceedings of the Fifteenth International Workshop on Data Warehousing and OLAP*, DOLAP '12, pages 1–8, New York, NY, USA. ACM.

Kimball, R. (1996). *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses.* Wiley.

Lenz, H.-J. and Thalheim, B. (2009). A formal framework of aggregation for the olap-oltp model. *Journal of Universal Computer Science*, 15(1):273–303.

Leonhardi, B., Mitschang, B., Pulido, R., Sieb, C., and Wurst, M. (2010). Augmenting olap exploration with dynamic advanced analytics. In *13th International Conference on Extending Database Technology (EDBT 2010)*.

Mahboubi, H., Ralaivao, J.-C., Loudcher, S., Boussaïd, O., Bentayeb, F., Darmont, J., et al. (2009). X-wacoda: an xml-based approach for warehousing and analyzing complex data. *Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction*, pages 38–54.

Messaoud, R. B., Boussaid, O., and Rabaséda, S. (2004). A new olap aggregation based on the ahc technique. In *DOLAP 2004, ACM Seventh International Workshop on Data Warehousing and OLAP*, pages 65–72.

Miquel, M., Bdard, Y., and Brisebois, A. (2002a). Conception d'entrepts de donnes gospatiales partir de sources htrognes. exemple d'application en foresterie. *Ingnieries des Systmes d'information*, 7(3):89–111.

Miquel, M., Bédard, Y., Brisebois, A., Pouliot, J., Marchand, P., and Brodeur, J. (2002b). Modeling multidimensional spatio-temporal data werehouses in a context of evolving specifications. *International Archives Of Photogrammetry Remote Sensing And Spatial Information Sciences*, 34(4):142–147.

Nguyen, T. B. and Tjoa, A. M. (2000). An object oriented multidimensional data model for olap. In *In Proc. of 1st Int. Conf. on Web-Age Information Management (WAIM), number 1846 in LNCS*, pages 69–82. Springer.

Phipps, C. and Davis, K. C. (2002). Automating data warehouse conceptual schema design and evaluation. In *Proceedings of the 4th International Workshop on Design and Management of Data Warehouses (DMDW)*, volume 2.

Romero, O. and Abello, A. (2009). A survey of multidimensional modeling methodologies. *International Journal of Data Warehousing and Mining*, 5(2):1–23.

Romero, O. and Abello, A. (2010). Automatic validation of requirements to support multidimensional design. *Data and Knowledge Engineering*, 69:917–942.

Sautot, L., Bimonte, S., Journaux, L., and Faivre, B. (2014a). A methodology and tool for rapid prototyping of data warehouses using data mining: Application to birds biodiversity. In *Proceedings of 4th International Conference on Model & Data Engineering (MEDI)*. In Press.

Sautot, L., Faivre, B., Journaux, L., and Molin, P. (2014b). The hierarchical agglomerative clustering with gower index: a methodology for automatic design of olap cube in ecological data processing context. *Ecological Informatics*. In Press.